Online Resource 3 Explanation of Partitioning around Medoids (PAM) methodology

Partitioning Around Medoids (PAM; Kaufman & Rousseeuw, 1990) was used for the clustering in this paper, using the implementation provided by the R package 'cluster'. PAM is a means to achieve *k*-medoid clustering. This is similar to *k*-means clustering, where *n* datapoints are partitioned into *k* clusters, however datapoints are used as the cluster centres rather than arbitrary points. The use of datapoints as cluster centres relaxes the need for a continuous distance metric, and allows the clustering of categorical data if a dissimilarity can be computed between any given points. PAM works roughly as follows:

- Given a partioning of the points around *k* medoids, a cost can be computed by summing the dissimilarity between each point and its medoid.
- An initial random set of *k* points are chosen as medoids.
- For each cluster, for each non-medoid point, try swapping the point to be the medoid, and recomputed cluster assignments and overall cost. If this does not decrease the cost, undo the swap
- The algorithm terminates when no decrease in cost is found

As with *k*-means, it is necessary to select a value for *k* before carrying out the clustering. The *silhouette* of a point *i* measures how well it matches the cluster it is assigned to compared to the closest other cluster, and is defined as (Rousseeuw, 1987):

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

This ranges from -1 to 1, with higher numbers indicating better assignments. The average silhouette width of a complete partitioning gives an indication of the quality of the clustering, and hence can indicate better or worse choices for *k*.

Other methods such as Level Cut Clustering could also be considered (Anderlucci & Hennig 2014), however it was felt that this was unlikely to make a significant difference.

Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics***20**: 53–65. doi:10.1016/0377-0427(87)90125-7.